Facultad de Informática
Grado en Ingeniería Informática

Lógica

PARTE 3: DEMOSTRACIÓN AUTOMÁTICA

Tema 12: Interpretaciones de Herbrand

Profesor: Javier Bajo
jbajo@fi.upm.es

Introducción.

❑ Componentes

Parte 1
Lógica Proposicional

Parte 2
Lógica de Primer Orden

Parte 3
Demostración Automática

Parte 4
Resolución

Introducción.

- $T[P_1, P_2, \ldots, P_n]$ |- C correcta sii $P_1 \wedge P_2 \wedge \ldots \wedge P_n \wedge \neg C$ insatisfacible
- Una fórmula *A* es insatisfacible sii ***no existe ninguna*** interpretación *I* tal que *I(A) = V*, o lo que es lo mismo
- *A* es insatisfacible sii todas las interpretaciones de la fórmula la evalúan como falsa

- Si *A* es una *fórmula proposicional* ➔ el número de interpretaciones a analizar sería finito: 2^n, siendo *n* el número de proposiciones diferentes que la integran
- Si *A* es una *fórmula de primer orden* ➔ el número de interpretaciones a analizar sería infinito y no numerable!!!

Introducción.

- Podríamos **simplificar** este problema si identificamos **algún subconjunto especial de interpretaciones de *A***
  - cuyo número sea menor (finito o infinito numerable como mucho)
  - cuyo análisis fuese suficiente para determinar la insatisfacibilidad de *A* (si *A* es insatisfacible para esas interpretaciones también lo es para cualquier otra)

- Esas interpretaciones existen para cualquier fórmula y se llaman **interpretaciones Herbrand**

Universo de Herbrand

- ***Universo de Herbrand de una fórmula A (H):*** Determina el dominio de interpretación de *A* en las interpretaciones Herbrand
- Está formado por todos los términos que se pueden definir con todas las constantes y funciones de *A* aplicadas de todas las formas posibles.
- Formalmente:
  - Sea *C* el conjunto de símbolos de constante de *A*
  - Sea *F* el conjunto de símbolos de función de *A*

$$H_0 = \begin{cases} C & \text{si } C \text{ no es vacío} \\ \{a\} & \text{si } C \text{ es vacío} \end{cases}$$

$$H_i = \{f(t_1,\ldots,t_n) \,/\, t_j \in (H_{i-1} \cup \ldots \cup H_0),\ f/n \in F\}$$

$$H = H_0 \cup H_1 \cup \ldots \cup H_i \cup \ldots \quad \text{es el universo de Herbrand de } A$$

Universo de Herbrand

Ejemplos:

$A = \{p(x), q(y)\}$

- $H_0 = \{a\}$
- $H_1 = H2 = \ldots = \varnothing$
- $H = \{a\}$

$A = \{p(a), q(y) \vee \neg r(b,f(x))\}$

- $H_0 = \{a, b\}$
- $H_1 = \{f(a), f(b)\}$
- $H_2 = \{f(a), f(f(a)), f(b), f(f(b))\}$
- …
- $H = \{a, b, f(a), f(b), f(f(a)), f(f(b)), \ldots\} = \{ a, b, f^n(a), f^n(b)\ \forall n \geq 1\}$

Base de Herbrand

- ***Átomo básico***: Se obtiene aplicando un símbolo de predicado de la fórmula a términos de su universo de Herbrand
- ***Base de Herbrand (BH)*** de una fórmula *A*: Conjunto de todos los átomos básicos de *A*
- Contiene todos los predicados de *A* aplicados a todos los términos de *H* de todas las formas posibles.
- Formalmente:
  - Sea *P* el conjunto de símbolos de predicado de *A*

$$BH = \{p(t_1,\ldots t_n) \;/\; t_j \in H,\; p/n \in P\}$$

Base de Herbrand

- $A = \{p(x), q(y)\}$
  - $H = \{a\}$
  - $BH = \{p(a), q(a)\}$

- $A = \{p(a), q(y) \vee \neg p(f(x))\}$
  - $H = \{a, f(a), f(f(a)), \ldots\} = \{ a, f^n(a)\ \forall n \geq 1\}$
  - $BH = \{p(a), p(f(a)), \ldots, p(f^n(a)), \ldots, q(a), q(f(a)), \ldots, q(f^n(a)), \ldots\}$
  - $BH = \{p(t), q(t)\ \forall t \in H\}$

- $A = \{p(a), q(y) \vee \neg r(b,f(x))\}$
  - $H = \{a, b, f(a), f(b), f(f(a)), f(f(b)), \ldots\} = \{a, b, f^n(a), f^n(b)\ \forall n \geq 1\}$
  - $BH = \{p(a), p(f(a)), \ldots, p(f^n(a)), \ldots, p(b), p(f(b)), \ldots, p(f^n(b)), \ldots, q(a), q(f(a)), \ldots, q(f^n(a)), \ldots, q(b), q(f(b)), \ldots, q(f^n(b)), \ldots, r(a,a), r(a,b), \ldots, r(a, f^n(a)), \ldots, r(b,a), \ldots, r(f^n(b), f^m(a)), \ldots\}$
  - $BH = \{p(t), q(t), r(t, t')\ \forall t,t' \in H\}$

Interpretación de Herbrand

- Una ***interpretación Herbrand (IH)*** de una fórmula *A* es una interpretación sobre *H* tal que:
  - Cada constante $a \in C$ se asocia consigo misma: $IH(a) = a$
  - Cada símbolo de función $f/n \in F$ se asocia con la función f_H/n
    - $f_H: H^n \Rightarrow H$
    - $IH(f(t_1, t_2,\ldots,t_n)) = f_H(IH(t_1), IH(t_2),\ldots, IH(t_n)) = f_H(t_1, t_2,\ldots, t_n) = t$
  - Cada símbolo de predicado $p/n \in P$ se asocia con el predicado p_H/n
    - $P_H: H^n \Rightarrow \{V, F\}$
    - $IH(P(t_1, \ldots, t_n)) = P_H(IH(t_1), \ldots, IH(t_n)) = P_H(t_1, t_2,\ldots, t_n) = V$ (o F)
  - Cada átomo básico de BH tiene un valor de verdad

  [C es el conjunto de constantes, F el conjunto de símbolos de función y P el conjunto de símbolos de predicado del LPO de A]

Interpretación de Herbrand

- Una interpretación Herbrand se puede representar como el conjunto de los átomos básicos de la Base de Herbrand,
  - afirmados si se interpretan como verdaderos
  - y negados si se interpretan como falsos

- Formalmente:

$$BH = \{A_1, A_2, A_3, \ldots\}$$

$$IH = \{A_1, \neg A_2, \neg A_3, \ldots\} \text{ si } IH(A_1) = V, IH(A_2) = F, IH(A_3) = F, \ldots$$

Interpretación de Herbrand

- $A = \{p(x), q(y)\}$
  - $H = \{a\}$, $BH = \{p(a), q(a)\}$

  Hay 4 interpretaciones de Herbrand:
  - $I_1 = \{p(a), q(a)\}$
  - $I_2 = \{p(a), \neg q(a)\}$
  - $I_3 = \{\neg p(a), q(a)\}$
  - $I_4 = \{\neg p(a), \neg q(a)\}$

- $A = \{p(a), q(y) \vee \neg p(f(x))\}$
  - $H = \{a, f(a), f(f(a)), \ldots\} = \{a, f^n(a)\ \forall n \geq 1\}$
  - $BH = \{p(a), p(f(a)), \ldots, p(f^n(a)), q(a), q(f(a)), \ldots, q(f^n(a))\} = \{p(t), q(t)\ \forall t \in H\}$

  Hay infinitas interpretaciones Herbrand, algunas de ellas son:
  - $I_1 = \{p(t), q(t)\ \forall t \in H\}$
  - $I_2 = \{p(a), \neg p(t)\ \forall t \in H - \{a\}, q(t')\ \forall t' \in H\}$
  - $I_3 = \{p(t), \neg q(t)\ \forall t \in H\}$
  - …………

Interpretación de Herbrand

¿Cómo saber el valor de verdad que una interpretación Herbrand asigna a una fórmula?

- *<u>Definición</u>:* ***Instancia básica*** de una cláusula de un fórmula en forma clausular es el resultado de sustituir las variables de la cláusula por términos de su universo de Herbrand

- Una interpretación Herbrand permite dar valor de verdad a una fórmula en forma clausular a partir del valor de verdad de las instancias básicas:
  - asignando valor de verdad a todas las instancias básicas (puesto que las variables están cuantificadas universalmente) de todas las cláusulas (puesto que están en conjunción)

Interpretación de Herbrand

- Ejemplo: $A = \{p(a), q(b) \vee \neg p(x)\}$
  - $H = \{a, b\}$, $BH = \{p(a), p(b), q(a), q(b)\}$
  - $IH = \{p(a), \neg p(b), q(a), \neg q(b)\}$
  - Instancias básicas de p(a): *p(a)* ya es una instancia básica, $IH(p(a)) = V$, por lo que **la primera cláusula de *A* es verdadera**
  - Instancias básicas de $q(b) \vee \neg p(x)$: $IH(q(b) \vee \neg p(a)) = F$ y $IH(q(b) \vee \neg p(b)) = V$, por lo que **la segunda cláusula de *A* es falsa**
  - Como *A* es la conjunción de ambas cláusulas entonces ***A* es falsa** para esta interpretación Herbrand

Interpretación de Herbrand

- *Definición*: Dada una interpretación I sobre un dominio D, una ***interpretación Herbrand correspondiente a I*** es una IH que satisface la siguiente condición:
  - A cada $t \in H$ le corresponde $I(t) = d \in D$
  - A cada átomo básico $p(t_1, \ldots, t_n) \in BH$ le corresponde el mismo valor de verdad que el que I asigna a $p(I(t_1), \ldots, I(t_n))$

- Puede haber varias IH correspondientes a la misma I:
  - Cuando la fórmula no tiene constantes, I no interpreta H_0 y por tanto la interpretación de $a \in H_0$ es arbitraria

- Ejemplo:
  - A = {p(x)}, D = {1, 2}, p(x): *x* es par
  - H = {a}, BH = {p(a)}
  - IH correspondiente a I: I(a) = 1, IH = {¬p(a)}
  - Otra IH correspondiente a I: I(a) = 2, IH = {p(a)}

Interpretación de Herbrand

- *Lema*: Si una interpretación I sobre un dominio D satisface una fórmula F, entonces cualquiera de las interpretaciones Herbrand correspondientes a I también satisfacen F

- *Teorema*: **Una fórmula F es insatisfacible sii F es falsa para todas sus interpretaciones Herbrand**

$\Rightarrow$ 1. F es insatisfacible

2. Todas las interpretaciones sobre todos los dominios la hacen falsa

3. En particular, todas las interpretaciones sobre el dominio H la hacen falsa

$\Leftarrow$ 1. F es falsa para todas sus interpretaciones Herbrand

2. Supongamos que F no es insatisfacible, entonces existiría al menos una interpretación I sobre algún dominio D que la satisfaría

3. Entonces las interpretaciones Herbrand correspondientes a esa I también satisfarían a F (por el lema anterior)

4. F sería verdadera para algunas de sus interpretaciones Herbrand (contradicción con 1)

5. F es insatisfacible

Interpretación de Herbrand

Por tanto, para estudiar la insatisfacibilidad de un fórmula basta con estudiar las interpretaciones Herbrand de su forma clausular

En la práctica: *Evaluación de una fórmula F en forma clausular*

- Para cada interpretación Herbrand de *F* :
  1) Definir las instancias básicas de las cláusulas
  2) Asignar valor de verdad a cada instancia según la interpretación
  3) Si alguna instancia básica se hace falsa entonces *F* es falsa
  4) Si ninguna instancia básica de ninguna cláusula se hace falsa entonces *F* es verdadera

- Si alguna interpretación Herbrand evalúa *F* como verdadera entonces *F* es satisfacible

- Si ninguna interpretación Herbrand evalúa *F* como verdadera entonces *F* es insatisfacible

Interpretación de Herbrand

- $F = \{p(x), q(y)\}$
  - $H = \{a\}$, $BH = \{p(a), q(a)\}$

  Hay 4 interpretaciones Herbrand:
  - $I_1 = \{p(a), q(a)\}$
  - $I_2 = \{p(a), \neg q(a)\}$
  - $I_3 = \{\neg p(a), q(a)\}$
  - $I_4 = \{\neg p(a), \neg q(a)\}$

  Instancias básicas de las cláusulas de F: $\{p(a), q(a)\}$
  - I_1 hace verdad todas las instancias, por lo que I_1 es un modelo de F
  - I_2, I_3 e I_4 falsifican una o dos de las instancias, así que todas ellas son contramodelos de F

  Como F tiene al menos un modelo (I_1) ➪ F es satisfacible

Interpretación de Herbrand

- F = {p(a), q(x) $\vee$ $\neg$p(f(x))}

  - H = {a, f(a), f(f(a)), …} = {a, f^n(a) $\forall$n$\geq$1}
  - BH = {p(a), p(f(a)),…, p(f^n(a)), q(a), q(f(a)),…, q(f^n(a))} = {p(t), q(t) $\forall$t $\in$ H}

  Hay infinitas interpretaciones Herbrand para *F*, algunas de ellas son:
  - I_1 = {p(t), q(t) $\forall$t $\in$ H}
  - I_2 = {$\neg$p(a), p(t) $\forall$t $\in$ H – {a}, q(t') $\forall$t' $\in$ H}
  - I_3 = {p(t), $\neg$q(t) $\forall$t $\in$ H}

  Instancias básicas:
  - De p(a): p(a) ya es una instancia básica
  - De q(x) $\vee$ $\neg$p(f(x)): q(a) $\vee$ $\neg$p(f(a)), q(f(a)) $\vee$ $\neg$p(f(f(a))), q(f(f(a))) $\vee$ $\neg$p(f(f(f(a)))),…

  I_1 hace verdad todas las instancias y por tanto es un modelo de *F*
  I_2 falsifica *p(a)*, por lo que *F* es falsa e I_2 es un contramodelo de *F*
  I_3 falsifica todas las instancias de la cláusula q(y) $\vee$ $\neg$p(f(x)), por lo que *F* es falsa e I_3 es un contramodelo de *F*

  Por tanto, F es satisfacible

Interpretación de Herbrand

- $F = \{p(y),\ q(a) \vee \neg p(f(x)),\ \neg q(x)\}$
  - $H = \{a, f(a), f(f(a)), \ldots\} = \{a, f^n(a)\ \forall n \geq 1\}$
  - $BH = \{p(a), p(f(a)),\ldots, p(f^n(a)), q(a), q(f(a)),\ldots, q(f^n(a))\} = \{p(t), q(t)\ \forall t \in H\}$

  Hay infinitas interpretaciones Herbrand para *F*, algunas de ellas son:
  - $I_1 = \{p(t), q(t)\ \forall t \in H\}$
  - $I_2 = \{q(a), \neg q(t)\ \forall t \in H - \{a\}, p(t')\ \forall t' \in H\}$
  - $I_3 = \{p(t), \neg q(t)\ \forall t \in H\}$

  Instancias básicas:
  - De $p(y)$: $p(a)$, $p(f(a))$, $p(f(f(a)))$, …
  - De $q(a) \vee \neg p(f(x))$: $q(a) \vee \neg p(f(a))$, $q(a) \vee \neg p(f(f(a)))$, $q(a) \vee \neg p(f(f(f(a))))$, …
  - De $\neg q(x)$: $\neg q(a)$, $\neg q(f(a))$, $\neg q(f(f(a)))$, …

I_1 falsifica todas las instancias de la cláusula $\neg q(x)$, así que I_1 es un contramodelo de *F*

I_2 falsifica la instancia $q(a)$ de la cláusula $\neg q(x)$, así que I_2 es un contramodelo de *F*

I_3 falsifica todas las instancias de la cláusula $q(a) \vee \neg p(f(x))$, así que I_3 es un contramodelo de *F*

De hecho, todas las posibles interpretaciones Herbrand de F son contramodelos ………

Interpretación de Herbrand

- $F = \{p(y),\ q(a) \vee \neg p(f(x)),\ \neg q(x)\}$

Instancias básicas:

De p(y):	De q(a) ∨ ¬p(f(x)):	De ¬q(x):
$p(a)$	$\boxed{q(a) \vee \neg p(f(a))}$	$\boxed{\neg q(a)}$
$\boxed{p(f(a))}$	$q(a) \vee \neg p(f(f(a)))$	$\neg q(f(a))$
$p(f(f(a)))$	$q(a) \vee \neg p(f(f(f(a))))$	$\neg q(f(f(a)))$
….	….	….

- Cualquier interpretación Herbrand que evaluase a *F* como verdadera, debería hacer verdad a <u>todas y cada una</u> de las instancias básicas anteriores

- Pero eso implica que, en particular, debería hacer verdad a las instancias: $p(f(a))$, $q(a) \vee \neg p(f(a))$ y $\neg q(a)$, y eso no es posible!

- Por tanto, no hay interpretaciones Herbrand que hagan verdad a *F* ⇨ *F* es insatisfacible

Teorema de Herbrand

Teorema de Herbrand: Un conjunto de cláusulas C es insatisfacible sii existe un conjunto finito de instancias básicas de cláusulas de C que es insatisfacible

- El teorema sugiere un procedimiento de refutación:
  - Dado un conjunto de cláusulas C insatisfacible, si se define un procedimiento mecánico que genere incrementalmente conjuntos de instancias básicas S_1, S_2, …, S_n, … de cláusulas de C y que analice de forma sucesiva la insatisfacibilidad de S_1, S_2, …, S_n, … entonces, como garantiza el teorema de Herbrand, este procedimiento puede detectar un k tal que S_k es insatisfacible.
- Implementación del teorema de Herbrand:
  - Método de Gilmore (1960)
  - Método de Davis-Putnam (1960)
  - Método de Resolución de Robinson (1965)

Estos métodos difieren en la forma de analizar la insatisfacibilidad de un conjunto de instancias básicas